Advancing the State-of-the-Art in Intelligent Systems: Scientific Rigor in Our Methods of Experimentation

Dennis K. Leedom, Ph.D.
Evidence Based Research, Inc.
Vienna, Virginia 22182

18 Sept 03

CCRP

1

Report Documentation Page

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
SEP 2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Advancing the State-of-the-Art in Intelligent Systems: Scientific Rigor in Our Methods of Experimentation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Evidence Based Research, Inc.,Vienna,VA,22182	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
PerMIS?03, Performance Metrics for Intelligent Systems, 16-18 Sep 2003, Gaithersburg, MD

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**22**	

Good morning.

What I would like to do this morning is to provide you with a brief overview of a recent guide published by the Office of the Assistant Secretary of Defense for Network Integration and Information, entitled the Code of Best Practice for Experimentation. The need for such a guide was envisioned a year or so ago by Defense officials who were concerned about the state of military experimentation within DoD –the manner in which it is planned, the manner in which it is executed, and the manner in which its findings contribute to DoD's overall program of military transformation. In short, this concern focused on the relative lack of scientific rigor inherent in many military experiments over the past decade or so.

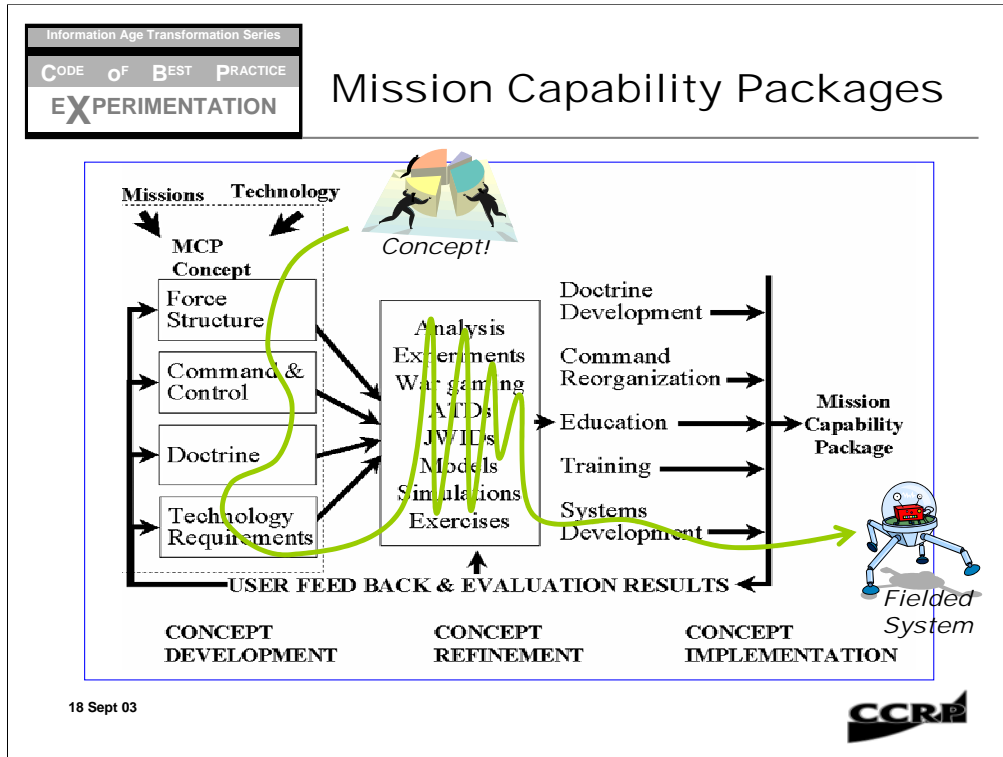As seen here, experimentation is considered the lynchpin in DoD's strategy for transformation. Without a properly focused, well-balanced, rigorously designed, and expertly conducted program of experimentation, the DoD will not be able to take full advantage of the opportunities that Information Age concepts and technologies offer.

The relevance of this topic to this NIST workshop should be obvious. Intelligent systems—for example, autonomous robotic vehicles, autonomous sensors, etc.—will likely play a significant role in future military operations. But unless those systems are properly designed and integrated into future concepts of operation, our investments in R&D will show little real payoff. To achieve this goal, most of your programs will involve experimentation as a necessary element of research and development. It is hoped that some of the lessons learned and captured in this new DOD code of best practice will help you designing experiments that are both scientifically grounded and operationally meaningful.

CODE **O**F **B**EST **P**RACTICE
E**X**PERIMENTATION

Mission Capability Packages

Missions   Technology

MCP Concept

Force Structure

Command & Control

Doctrine

Technology Requirements

Concept!

Analysis
Experiments
War gaming
ATDs
JWIDs
Models
Simulations
Exercises

Doctrine Development

Command Reorganization

Education

Training

Systems Development

Mission Capability Package

Fielded System

USER FEED BACK & EVALUATION RESULTS

CONCEPT DEVELOPMENT

CONCEPT REFINEMENT

CONCEPT IMPLEMENTATION

18 Sept 03

CCRP

Before addressing the specific steps involved in planning and executing an experiment, it is useful to place experimentation in a programmatic context. In this regard, DoD has adopted the notion of a "mission capability package" to denote the various elements that must be combined when developing and fielding a new piece of technology. As shown here, equal consideration must be given during concept development to (1) how the technology will fit into the overall force structure, (2) how the technology will be controlled and operated, (3) how the technology is envisioned to contribute to force effectiveness, and (4) what are the major technological hurdles to be overcome. Since many of these questions and their interactions are unknown at first, most development programs will go through a phase of experimentation and analysis to refine these concepts into a meaningful whole. What should emerge at the point of implementation is a cohesive mission capability package that addresses (1) the required changes to military doctrine brought about by the new technology, (2) the required modifications or improvements in command and control needed by the new technology, (3) the education and training required of the personnel expected to operate and employ the new technology, and—finally—the systems development required to produce a fieldable system.
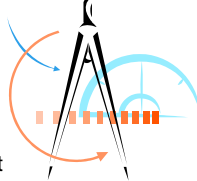
If any of these components are missing—as is frequently the case with many new areas of technology—the resulting implementation will result in years of confusion and frustration as the receiving military units attempt to incorporate a new system into their concept of operation.

Thus we see that experimentation must address not only the technological aspects of a new system, but also the doctrinal, C2, training, organizational, and personnel aspects as well.

3

Types of Experiments

- **Discovery Experiments**
  - Involves introducing novel systems, concepts, organizational structures, technologies, and other elements into a setting where their use and interaction can be observed and classified
  - Objectives: Identify potential military utility, refine research focus
  - Issues: Typically lack the control needed to infer cause-and-effect
- **Hypothesis Testing Experiments**
  - Involves the testing of hypothesized effects of a new system, concept, technology, etc. on different levels of performance under controlled conditions
  - Objectives: Isolate and reliably measure specific cause-and-effect paths
  - Issues: Difficult to establish experimental controls in a field setting
- **Demonstration Experiments**
  - Involves the recreation and synthesis of well-established physical, informational, cognitive, and social effects to demonstrate improvement in military capability
  - Objectives: Display existing knowledge and achievements to decision makers
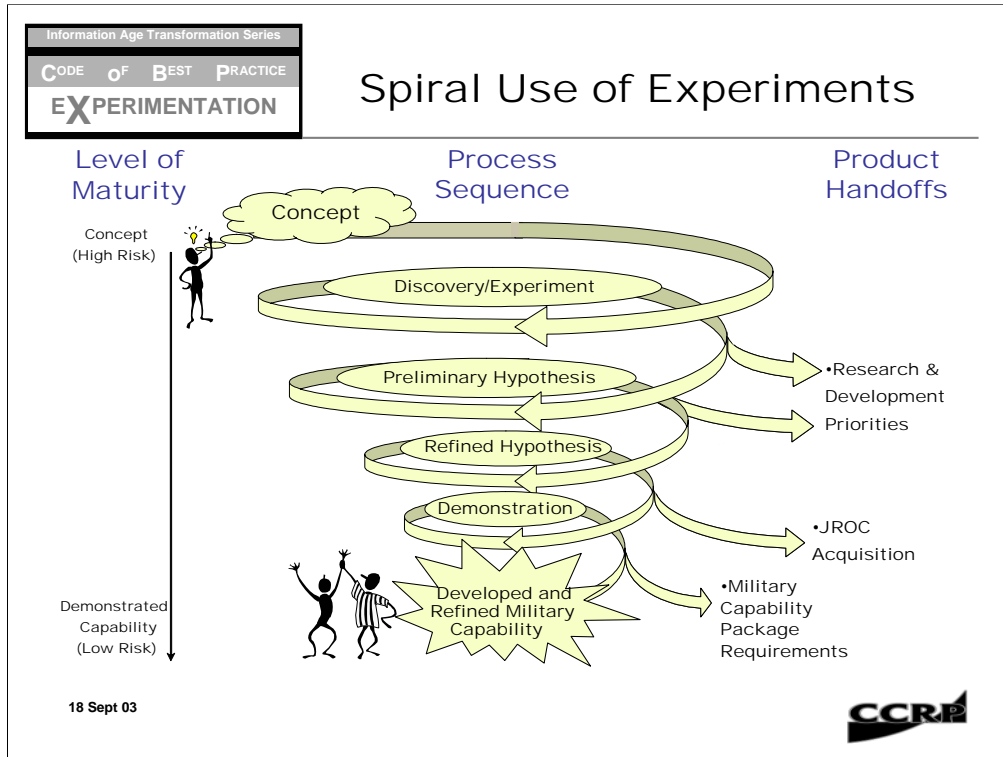 - Issues: Experimental context sometimes lacks operational relevancy / realism

18 Sept 03

Another point to consider is that not all experiments are created equal –or, at least not created for equal purposes. Shown here are the three major types of experiments that typically evolve during the course of a research and development program. Discovery experiments—considered early in the process—are typically employed to explore novel technologies or operational concepts. Such experiments have not yet matured to the point where system developers can make specific predictions as to the impact of the technology on military operations. At the same time, the lack of scientific rigor in discovery experiments often leaves little opportunity for discovering cause-and-effect.

Shown in the middle are hypothesis testing experiments that come after a technology program has matured to the point where specific predictions can be made regarding the anticipated impact on operational performance and effectiveness. It is at this point where the need for scientific control and rigor is the greatest. To sell the merits of a new technology, its developers must be able to generate convincing evidence of cause-and-effect –not merely that the new technology works in some functional way. Isolating specific pathways of cause-and-effect require good experimental controls – something that is often difficult to achieve in a field setting. At the same time, developers should consider such experiments as a valuable source of insight for refining system concepts and design.

Demonstration experiments come near the end of a development program when it is useful to provide evidence that the various components of the mission capability package have been properly considered and integrated into an effective military capability. Such evidence is used by senior decision makers to determine if the new technology (1) is sufficiently mature and (2) makes a sufficient contribution to warrant further investment and fielding. At issue typically in such experiments is the need for operational relevance and realism.

Spiral Use of Experiments

Information Age Transformation Series

CODE oF BEST PRACTICE
EXPERIMENTATION

Level of Maturity	Process Sequence	Product Handoffs

Concept (High Risk)

Concept

Discovery/Experiment

Preliminary Hypothesis

Refined Hypothesis

Demonstration

Developed and Refined Military Capability

Demonstrated Capability (Low Risk)

•Research & Development Priorities

•JROC Acquisition

•Military Capability Package Requirements

18 Sept 03

CCRP

There is no single, linear path that every program manager should follow in laying out a sequence of experiments. As depicted here, the process is often a spiral combination of experimentation, analysis, and concept refinement as program officials seek to translate technological advances into workable and useful operational systems.
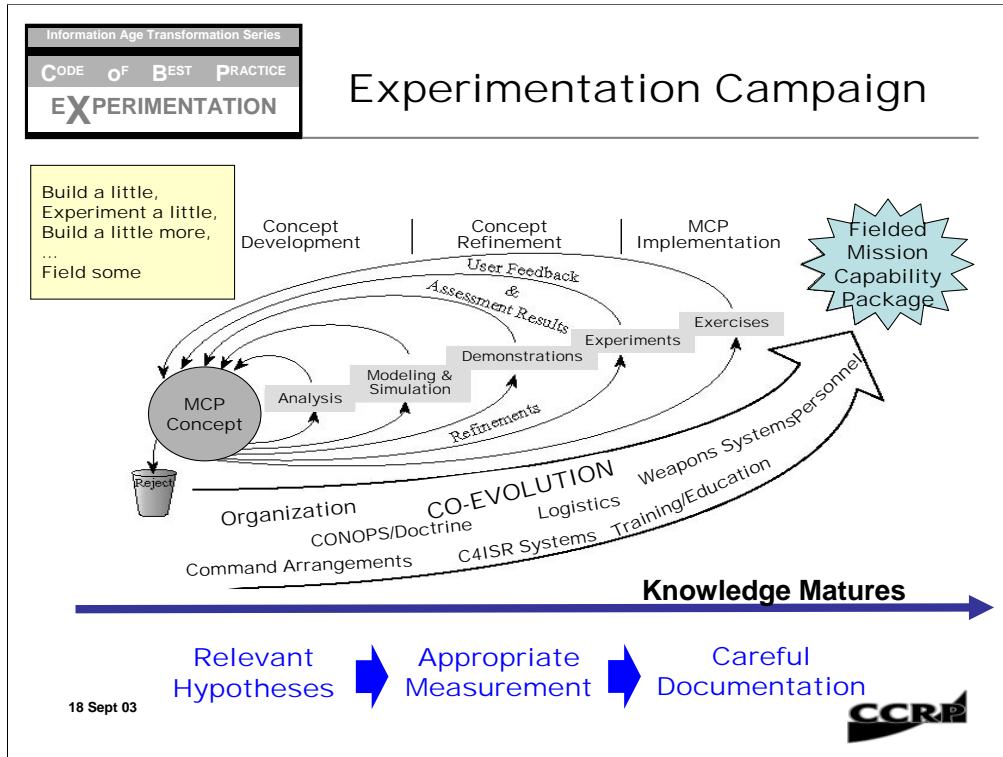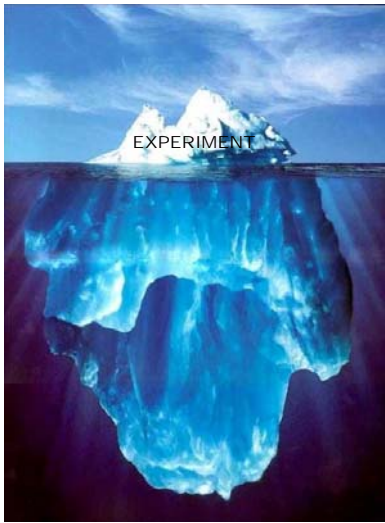
At the same time the basic technology is being refined and matured, there should also occur a maturation of the other elements that comprise the total mission capability package. The point here is that this is a co-evolution process –one in which each of these elements evolve simultaneously and in synchronization with one another. Unfortunately, a more common practice is for program managers to leave many of the non-technological issues unaddressed until a new technology is about ready for fielding. The relevance of this concern for experimentation is that this co-evolution process must be kept in mind when planning and executing various types of experiments throughout the development cycle.

Having discussed the programmatic context of experimentation, we now turn to the actual process of planning and executing a good experiment. Here, I ask the question, "Most of the effort is expended during the experiment …true or false?"

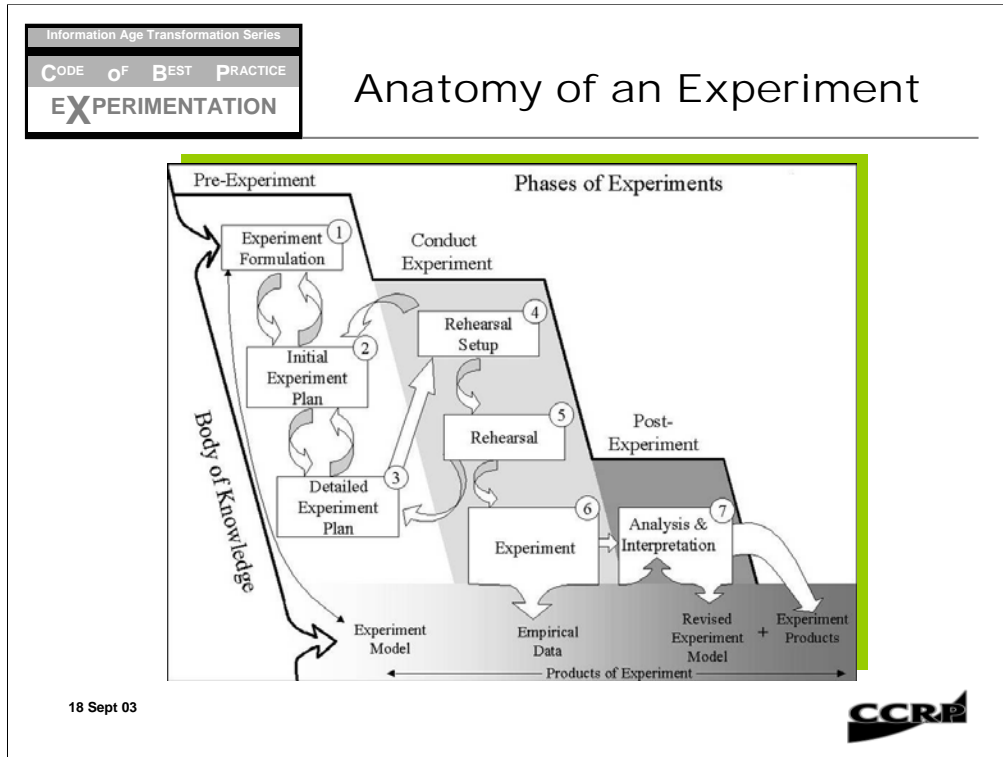While it might seem from the apparent level of activity exhibited in some development programs that most of the effort is expended during an experiment, the basic answer to this question should be "False!" In fact, the bulk of the effort in successful experiments is invested before the experiment itself is conducted. Additionally, substantial effort is also required after the experiment is conducted when the results are analyzed, understood, extrapolated, documented, and disseminated.

What should be obvious from a scientific perspective, however, is not always obvious to program managers who are often seen frantically assembling an experiment in a "just-in-time" manner, followed by a quick—and often incomplete—analysis of preliminary findings simply to meet bureaucratic deadlines. Unfortunately, the actual execution of experiments too often gets all of the publicity and attention from senior decision makers. Little attention is often given to the proper planning of a military experiment or to the proper analysis and interpretation of findings.

Thus, the picture of an iceberg—with its major portion hidden below the surface—is an apt description of how a program manager should approach the budgeting of time and resources for an experiment.

Anatomy of an Experiment

Information Age Transformation Series
CODE OF BEST PRACTICE
EXPERIMENTATION

Pre-Experiment — Phases of Experiments

1. Experiment Formulation
2. Initial Experiment Plan
3. Detailed Experiment Plan
4. Rehearsal Setup
5. Rehearsal
6. Experiment
7. Analysis & Interpretation

Conduct Experiment

Post-Experiment

Body of Knowledge

Experiment Model — Empirical Data — Revised Experiment Model + Experiment Products
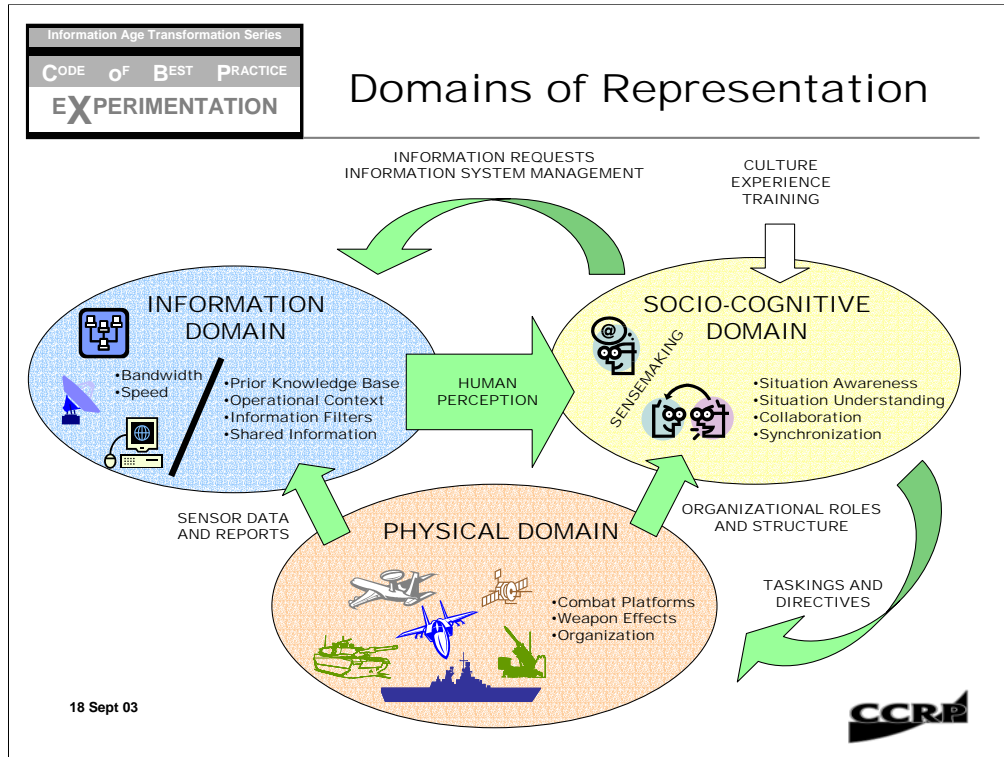
Products of Experiment

18 Sept 03

CCRP

If we were to break down this picture of an iceberg in more detail, we would see that the planning, execution, and exploitation of an experiment can be broken into several phases. Shown here is one of the illustrations from the Code of Best Practice book. The phases consist of (1) pre-experiment formulation and planning of experimental hypotheses, controls, and procedures; (2) the actual execution of the experiment that includes the important step of rehearsal; and (3) the analysis, interpretation, and documentation of the experimental findings.

Because of its central importance, I would like to spend the next few minutes talking about several issues that must be considered during the pre-experimental phase.
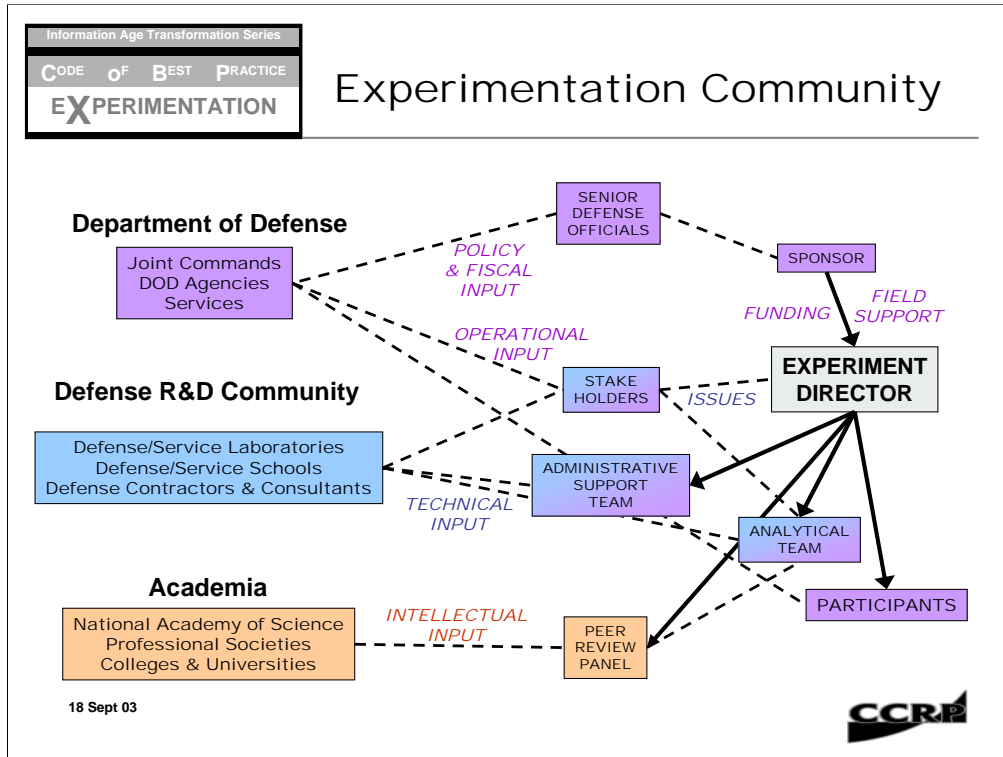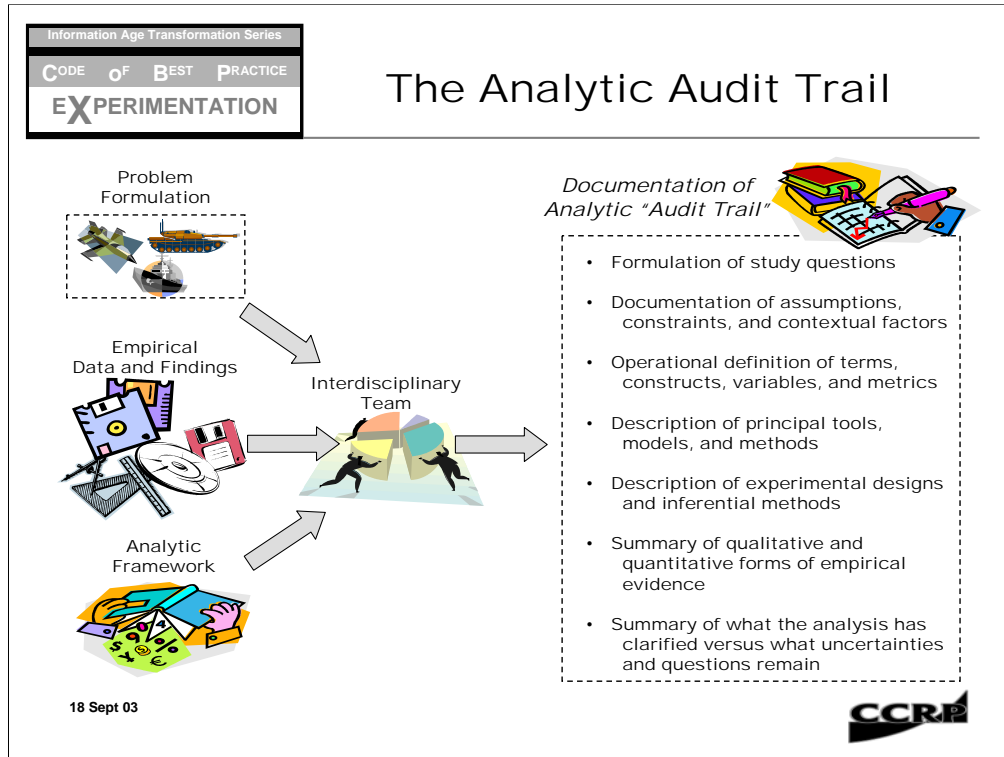
The first issue deals with proper identification of the experimentation community – that is, the community of stakeholders who have a vested interest in (1) what the experiment addresses, (2) how the experiment is conducted, and (3) what the experiment discovers in the way of empirical evidence and insights. As shown here, this community is often comprised of three elements. The first element are those Defense agencies, commands, and services who provide the program funding and operational support necessary to conduct the experiment. The second part of this community includes the laboratories, service schools, and defense contractors who provide technical input for shaping the experiment. The final element of this community are the academicians and professional societies that provide intellectual oversight to the experiment.

Together, these various stakeholders have vested interests in how the experiment is planned and executed. Thus, it is important for the experiment director to identify and include important members of this community early-on in the planning activities.

Domains of Representation

INFORMATION REQUESTS
INFORMATION SYSTEM MANAGEMENT

CULTURE
EXPERIENCE
TRAINING

INFORMATION DOMAIN

•Bandwidth
•Speed

•Prior Knowledge Base
•Operational Context
•Information Filters
•Shared Information

HUMAN PERCEPTION

SOCIO-COGNITIVE DOMAIN

SENSEMAKING

•Situation Awareness
•Situation Understanding
•Collaboration
•Synchronization

SENSOR DATA AND REPORTS

PHYSICAL DOMAIN

•Combat Platforms
•Weapon Effects
•Organization

ORGANIZATIONAL ROLES AND STRUCTURE

TASKINGS AND DIRECTIVES

18 Sept 03

CCRP

Looking at a typical experiment from a different perspective, we see here that most technologies emerging in today's Information Age involve three different domains of representation. The physical domain includes those materiel systems and their physical operation and impact on the battlefield. The information domain includes all of the concomitant knowledge bases, communication systems, and other information technology needed to command and control a specific technology on the battlefield. For those of you involved in the development of <u>intelligent</u> systems, consideration of this domain should be obvious. However, it is unlikely that many military systems will operate in a vacuum devoid of human intervention at some point. Thus, it is necessary to consider socio-cognitive domain that reflects the personnel and organizations that will maintain and operate the new technology on the battlefield.

As shown in this slide, each of these domains interact with the others. Thus, to arbitrarily isolate one domain from the others in an experiment is naïve and likely to produce a flawed experimental design. On the other hand, proper consideration of each domain in an experiment can be challenging because the relevant fields of expertise are likely to come from different academic or research backgrounds. As a result, many experiments reflect the failure of program managers to include each of these relevant fields of expertise in the planning team.

The Analytic Audit Trail

Information Age Transformation Series
CODE OF BEST PRACTICE
EXPERIMENTATION

Problem Formulation

Empirical Data and Findings

Interdisciplinary Team

Analytic Framework

Documentation of Analytic "Audit Trail"

- Formulation of study questions
- Documentation of assumptions, constraints, and contextual factors
- Operational definition of terms, constructs, variables, and metrics
- Description of principal tools, models, and methods
- Description of experimental designs and inferential methods
- Summary of qualitative and quantitative forms of empirical evidence
- Summary of what the analysis has clarified versus what uncertainties and questions remain

18 Sept 03

CCRP

This point brings us to the next issue to consider during the pre-experiment phase: proper construction of the analytic audit trail. The concept of an audit trail is important because of the need to anticipate—and consciously plan for—what types of evidence and insights are expected to be produced by an experiment. Here, it is necessary to work backwards from the anticipated findings to identify (1) the proper formulation of study questions, (2) the operational definition of critical performance measures and metrics that must be collected, (3) the types of statistical analyses and models that will be required to produce the needed insights, and (4) the types of experimental controls and scenarios that will be needed to provide an appropriate context for collecting the measures and metrics.

Too often, program managers approach experiments without a clear and concise set of study questions that are posed in the form of testable hypotheses. Too often is the case where data collection is defined more by convenience, rather than by what is relevant to the study issues. And too often is the case where the data analysis plan is considered only after-the-fact when the experiment has been completed and everyone is anxiously awaiting a publication of findings. Each of these situations are likely to produce little, if any, useful findings and to be wasteful of the time and resources expended on the experimental event.

To repeat the obvious, proper design of the analytic audit trail should proceed from the very beginning of the planning phase. To quote an old proverb, "If you don't know what you're looking for, you'll never find it!"
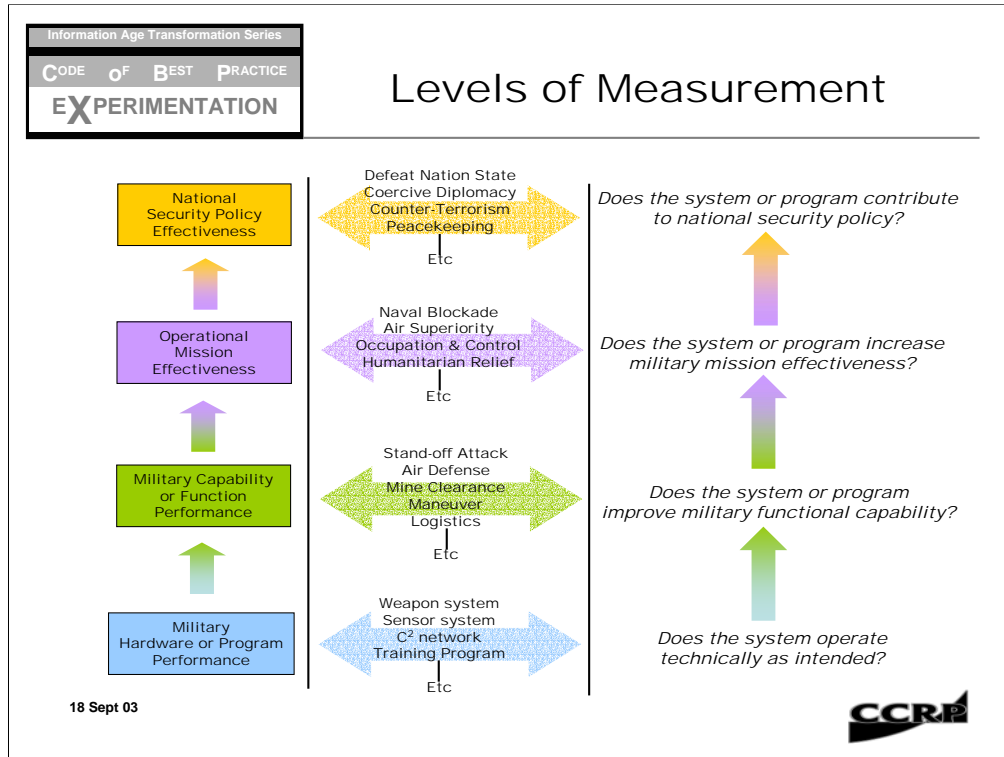
As part of building the analytic audit trail, experiment planners must consider different levels of system measurement that might be relevant to the study questions. Depicted here are various levels of system measurement, extending from basic hardware or software performance, up through the contribution that a technology might make to some functional capability, up through the impact of some improved function on operational mission effectiveness, finally up to the level of national security policy effectiveness. A common flaw seen in some experimental designs is the tendency of technology developers to attempt to jump right from hardware performance to impact on operational effectiveness or security policy effectiveness. In reality, there might be only a very loose correlation between these two levels of outcome measure. A better approach to experimentation is one that carefully identifies how a particular technology development might impact at each level. While this increases the amount of data that must be collected during an experiment, it is often necessary so that analysts can later build a complete story that links findings at one level of measurement up through the other levels of measurement.
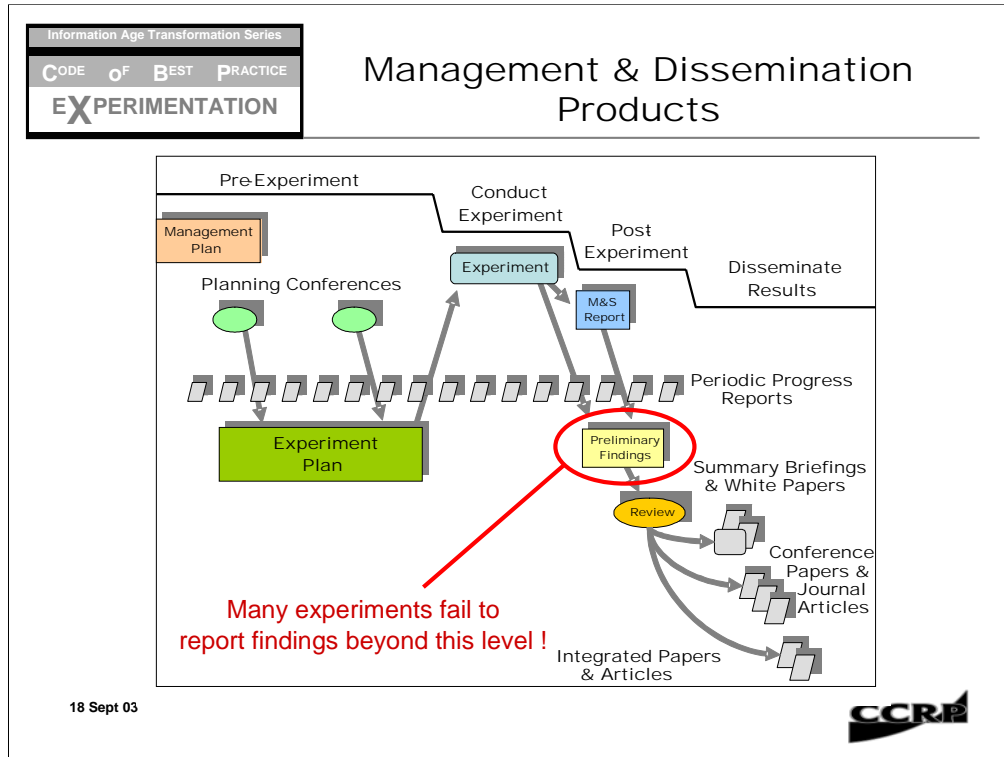
Finally, we come to the issue of documentation –an often overlooked aspect of planning and conducting an experiment. Depicted in this slide are the various types of documents required for good experimentation. Documentation begins in the pre-experiment phase so that each relevant stakeholder and contributor within the experimentation community will have a clear understanding of what will transpire during the experiment. For large experiments involving several military units, planning conferences are typically held months—if not years—in advance of the actual experimental event. The issuance of periodic progress reports will help to keep community members informed of critical changes in the design that might be necessary due to scheduling or resource conflicts.

Documentation is also critical after the completion of an experimental event. All too often, the main report that gets published is a "quick look" briefing that reports out preliminary observations from an experiment –usually to satisfy the political needs of the bureaucracy. Unfortunately, attention to experimental findings quickly diminishes after such a briefing as decision makers and their supporting staff turn to other projects. As a result, the normal production of scientific papers and in-depth analyses gets placed on the "back-burner" –thus denying the opportunity for the experiment to truly add to the body of scientific knowledge.

Proper, documentation of experimental findings is a prerequisite of good science, and it is up to the conscious volition of the program manager to insure that this important step gets accomplished.
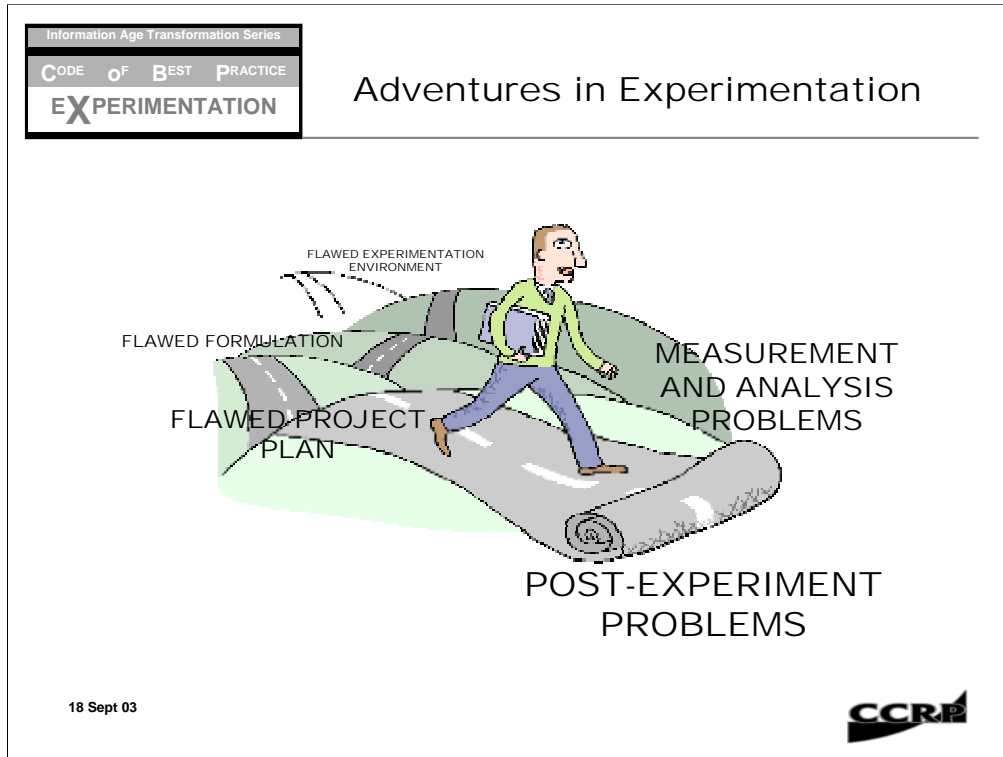
Having reviewed the basics of good experimentation as outlined in the Coed of Best Practice, I would like to spend the remaining few minutes highlighting the types of flaws most often seen in military experiments. Such problems fall into the several categories shown here, and each can significantly degrade or negate the return-on-investment from a given experiment.

Details of such flaws are covered in more detail in a separate chapter in the Code of Best Practice.

Flawed Experimentation Environment

1. **Piggy-backing research experiments onto military training exercises**
   – Training objectives will conflict with experimental control and measurement
   – There is little opportunity provided to fail –and, subsequently, to learn
2. **Introducing multiple experiments into a single venue**
   – Confusion will arise regarding factors, experimental variables, and outcome measures
3. **Partial implementation of a developmental technology**
   – Lack of logistics, C2 infrastructure, etc. will confound performance
4. **Advocacy experiments designed to showcase a specific initiative**
   – Lack of objective context makes it difficult to judge contributions
5. **Failure to permit a thinking, reactive adversary**
   – Situation does not allow identification of conceptual or technical weaknesses / loopholes
6. **Reliance on a single scenario**
   – Situation does not allow findings to be generalized to more relevant operational contexts
7. **Use of inappropriate or poorly trained test subjects**
   – Lack of experience and expertise with new concepts, technology, etc. will negate any potential benefits

18 Sept 03

Our list begins with problems commonly associated with a flawed experimentation environment. Perhaps the biggest problem seen in recent years is the temptation to piggy-back experiments onto scheduled training exercises –a strategy usually motivated by resource limitations. The problem with this approach is that experimentation objectives will inevitably conflict with training objectives –with training objectives usually being priority. Conflicts usually show up in the area of experimental controls and measurement opportunities. As a result, such experiments provide little or no opportunity for objective learning.

Other problems seen with lesser frequency include (1) attempts to introduce multiple experiments into a single event or venue, (2) experimentation with technologies that are not yet ready for prime-time, (3) distortion of experimental controls and findings by political advocates of a particular technology, (4) the failure to test technologies against a smart adversary, (5) reliance on a single scenario that precludes generalization of findings, and (6) the use of inappropriate or poorly trained test subjects –e.g., equipment operators.

Flawed Formulation

1. **Experimentation team lacks broad experience and relevant skills**
   – Leads to poor understanding of problem space under study
   – Often ignores the "soft" disciplines of cognitive/social psychology, organizational theory
2. **Lack of meaningful and testable hypotheses**
   – "If we do X, Y, and Z, then our system will perform successfully"
   – All human test subject groups are assumed to be equal
3. **Inadequate manipulation of the independent variable(s)**
   – Provides little or no basis for establishing causality
4. **Failure to control for human subjects and organizational variables**
   – Leads to serious confounding of experiment and the inability to attribute specific performance differences to the concept or technology under study
5. **Experiment lacks a baseline or comparative case**
   – Sponsors and stakeholders are left with only anecdotal evidence
6. **Experiment implements only part of the mission capability package**
   – "Missing elements" are often ignored because of cost constraints
   – Yet, this problem often leads to understated / negative performance improvements
7. **Experimentation campaign lacks explicit model of the problem**
   – Lack of clear vision of problem often leads to "let's try it and see what happens"
   – Analysts will have great difficulty organizing the experimentation "story"
   – Leads to lack of continuity from one experiment to the next

18 Sept 03

Problems can also arise with the formulation of the experimental design. One overarching problem seen in many experiments is the lack of relevant expertise on the planning team that leads to an inadequate consideration of one or more of the domains discussed earlier. Military experiments are also prone to lacking meaningful and testable hypotheses. Substituted here are vague statements such as "This technology will increase performance" or "Let's just show that this technology can work…"

Inadequate manipulation of independent variables reflects the fact that the experimental design did not provide for an adequate range of performance differences between different cases –a prerequisite for establishing causality. Other problems include (1) failure to control for human subjects and organizational variables that impact on performance, (2) the lack of a meaningful baseline to which comparisons can be drawn, (3) partial implementation of a mission capability package that lacks one or more of the essential elements discussed earlier, and (4) the lack of an explicit performance model needed to link findings at one level of system measurement to the other levels.

Flawed Project Plan

1. **Experiment is started from a narrow base, ignoring other projects**
   - Duplication / overlap of experiments is enormously wasteful of time and energy
   - Fails to capitalize on existing knowledge and lessons learned from other projects
   - Often results from a failure to expose research design to peer review
2. **Experiment lacks formal data collection and analysis plan**
   - Lack of plan leads to strategy of "let's capture everything" or "let's only collect what we can easily measure –not what's important"
   - Excessive burden is placed on analysis phase to make sense of a "pile of data"
3. **Failure to hold a rehearsal –i.e., last-minute set-up**
   - Rehearsals guard against the inevitable "unexpected" glitches in designs, procedures, and measurement methods
   - Often leads to failed experiments, wasted resources, and program delays
4. **Analysis of findings is rushed to satisfy program/budget decisions**
   - "Quick look" reports and "initial assessments" are often scheduled within days (or hours) of the completion of the experiment
   - The subsequent analysis of empirical findings often contradicts initial impressions of human observers and experiment controllers
5. **Failure to control for visitors**
   - Senior officers or experts can have disruptive impact on behavior of test subjects
   - Best accomplished on a separate day set aside for visitors or public observers
6. **Failure to debrief all participants**
   - Participants see the experiment from different perspectives –regardless of role
   - Too often, participants are allowed to "disappear" to other assignments before their insights can be effectively captured

18 Sept 03

CCRP

Problems associated with a flawed project plan are listed here. A principal error observed in many programs is the tendency to ignore research findings from prior studies and experiments. This type of "go it alone" attitude is both wasteful of resources and violates one of the basic elements of good science –building upon prior knowledge. As mentioned earlier, experimenters often rush to conduct an experiment without a good data collection and analysis plan –one built one a solid analytic audit trail. Such an approach leads to both (1) confusion over how to make use of data that is collected and (2) a failure to address critical study questions. The failure to conduct rehearsals—often done for the sake of expediency—ignores Murphy's Law and just flies in the face of common sense; yet, we often see problems emerge during an experiment that could have been easily corrected had the planners conducted a simple rehearsal.

As mentioned earlier, there is often a rush to publish preliminary findings in order to satisfy the political demands of senior decision makers. The problems with this approach is that subsequent, in-depth analysis of the experimental data frequently contradict earlier subjective observations –thus producing a misleading impression of a particular technology. Failure to control for visitors is a problem seen in those experiments attracting a lot of attention from the community. While visitors should be allowed to witness aspects of the experiment, they should not be allowed to unduly influence test participants by their presence –a problem particularly associated with flag rank visitors. Finally, there is often a tendency to not debrief all test participants –even those who play a secondary or support role. Often such individuals will provide a unique perspective on the experiment and offer insight into why the experiment produced novel or unexpected findings. Too often, such participants quickly disappear to other assignments before their insights can be effectively captures.

Measurement and Analysis Problems

Information Age Transformation Series

CODE **O**F **B**EST **P**RACTICE

E**X**PERIMENTATION

1. **Reliance upon the "happiness test" instead of empirical data**
   - Common approach taken when experiment is observed by senior officers
   - Ignores the fact that people commonly "see what they want to see"
2. **Focus only on what can be easily measured, not what is relevant**
   - Wastes data collection / analysis resources while yielding little insight
   - Reflects lack of an overall model of problem
3. **Inadequate opportunity / access for relevant observation**
   - This problem is most critical in areas where performance depends upon human decision making
   - Requires good observer training and rehearsal to make data collection unobtrusive
4. **Confusing measures of <u>performance</u> with measures of <u>effectiveness</u>**
   - Measures of performance deal with the functioning of the system or technology
   - Measures of effectiveness deal with impact on military capability and outcome
5. **Failure to capture and explain anomalous events and time periods**
   - Not all phases of an experiment go as planned
   - Separating performance data by event/time yields additional/more accurate insights
6. **Failure to properly select and train observers / controllers**
   - Trade-off between relying upon academic versus military observers
   - Training is necessary to sensitize and calibrate their observations
   - Recalibration might be needed mid-way through an experiment
   - Inter-coder reliability tests should be used when relying upon human judgment

18 Sept 03

CCRP

Measurement and analysis problems are listed here. Foremost on the list is the tendency of experimenters to rely upon the "happiness test" provided by senior flag officers, instead of the rigorous use of empirical performance measures. This is a particularly challenging problem for technologies and concepts that involve the information and socio-cognitive domains mentioned earlier –two areas where some effort must be expended to develop good operational performance measures. A corollary to this problem is the tendency to focus only on what can be easily measured, and not what is relevant to answering the study questions.

Other problems include (1) inadequate opportunity for placing observers and data collectors where they can collect meaningful data, (2) the confusion of measures of performance with measures of effectiveness, (3) the failure to capture and explain anomalous events and time periods during an experiment, and (4) the failure to adequately select and train data collectors and experiment controllers for their critical roles.

CODE **O**F **B**EST **P**RACTICE
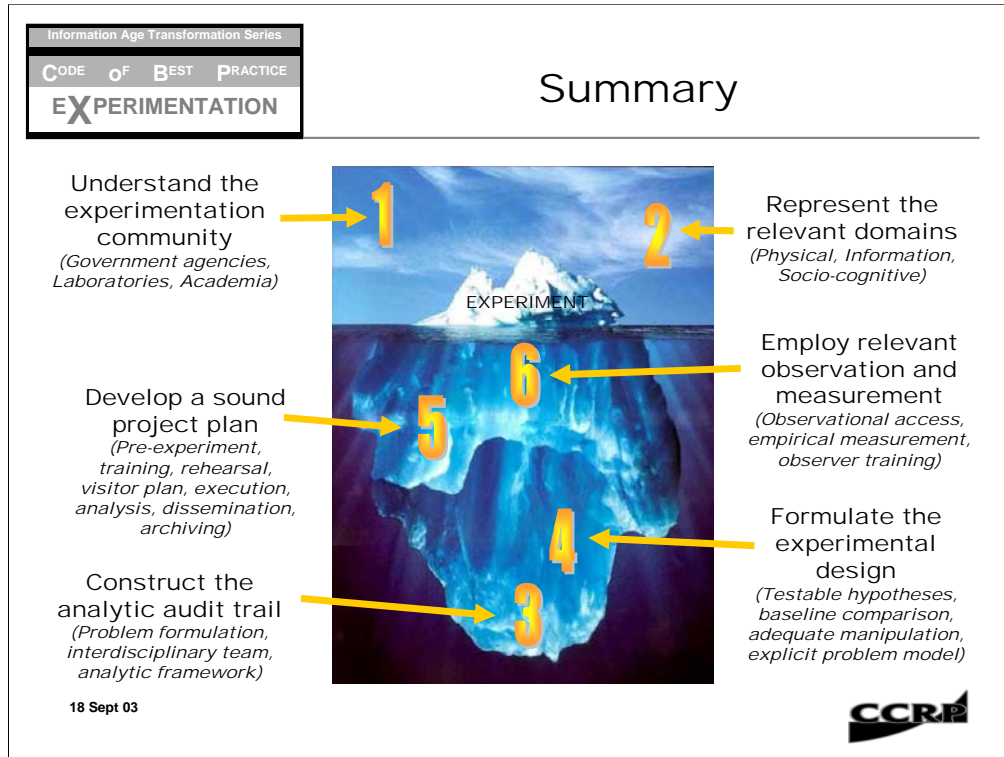E**X**PERIMENTATION

Post-Experiment Problems

1. **Failure to make experimentation data available to others**
    – Common strategy is to limit data access to core research team, but this denies opportunity for peer review and alternative interpretations
    – Occasionally done for reasons of security classification; however, more frequently, it is because of the fear of potential embarrassment
2. **Failure to archive experimentation materials**
    – Denies opportunity for revisiting data in the context of future problems
    – Obscures or masks artifacts of the experiment overtime, thus leading to future misinterpretation of the findings and insights
    – Wasteful of the time and resources invested in the experiment

18 Sept 03

CCRP

Finally, we list a number of post-experiment problems. Here, one of the biggest problems seen in many experiments is the failure of the program manager to make experimentation data available to others for reanalysis and reinterpretation. Occasionally this is done for reasons of security classification; however, more frequently, it is because of the fear of potential embarrassment over disappointing findings with a specific technology or system. A corollary to this problem is the failure to archive experimentation materials—e.g., experimental design, scenario, test probes, data collection methods, etc.—so that others can correctly interpret findings and their implications in other settings.

So, in summary, we see that a lot of careful thinking, planning, and coordination goes into a successful experiment –one not only based on scientific principles, but also one that effectively contributes to the needs of senior decision makers. As suggested by the iceberg analogy, much of this work occurs below the level of publicity and the attention of senior decision makers. Yet, each element shown here is absolutely critical to achieving good return-on-investment on the time and resources committed to an experiment.

These lessons learned have been derived from observing numerous military experiments over the past decade or so –many of which involved the testing of technologies related to intelligent systems, the focus of this workshop. These lessons, however, are timeless and apply to the very work you are contemplating over the next year.
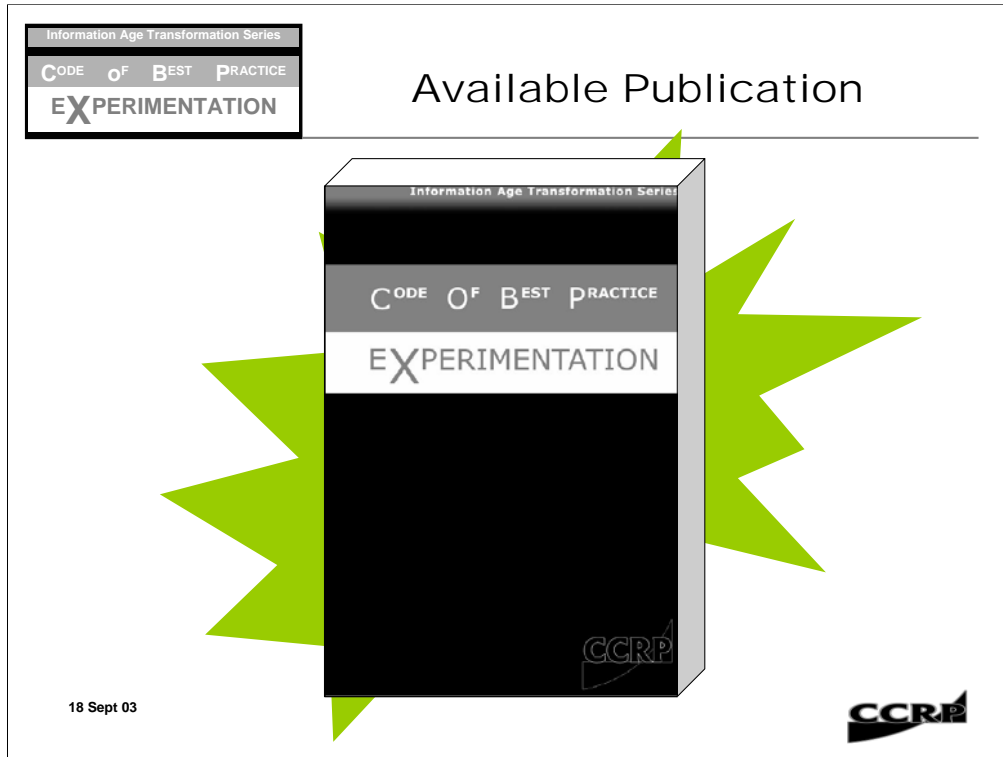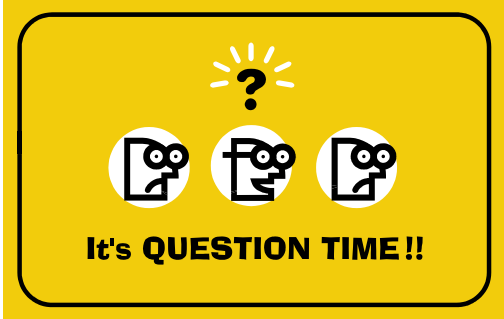
Let me conclude by noting that copies of this new guide are available by contacting the DoD Command and Control Research Program. This organization maintains a website at **www.dodccrp.org** where copies of the book can be ordered or downloaded in PDF format.

Thanks for your time and attention. I will now entertain a few questions from the audience.